Universal dependencies for Scottish Gaelic: syntax

Colin Batchelor

Royal Society of Chemistry, Cambridge, UK

19th August 2019

Overview

Motivation

- No validated syntactic parser for Scottish Gaelic.
- Contributing to a multi-language project enables us to concentrate on the data while other people concentrate on the tools (tokenisers, sentence-splitters, parsers, and so on.)

Scottish Gaelic

- Around 60 000 speakers.
- Separate written form from Irish since *ca.* the Jacobite rebellion.
- VSO (*cf.* Breton).
- Usual way of forming the present tense is *bi* + a verbal noun.
- Compound/inflected prepositions like other Celtic languages.
- Multiple ways of forming copular statements.
- Cosubordination and other unusual uses of coordinators *agus*/*is*/*'s*.
- Morphology: orthographically-marked lenition (except for l, n and r) and slenderisation.
- Resources: mainly `dasg.ac.uk`, *Am Faclair Beag* and other work by Michael Bauer, work by Kevin Scannell, Caoimhín Ó Donnaíle's resources at SMO and ARCOSG.

ARCOSG (Lamb *et al.* 2014–2016)

First machine-readable corpus of Scottish Gaelic.

- Over 80 000 tokens in 77 documents.
- Over 1300 utterances; around 1800 sentences (sentence/utterance boundaries not supplied).
- Categories: conversation, interview, sports commentary, narrative, news scripts, fiction, formal writing, popular writing.
- We also know the authors of the written text. (*cf.* `https://www.aclweb.org/anthology/P19-1339`)
- Identifying dialect takes a bit more digging.
- Tagged with parts of speech based on the PAROLE scheme.

Online here: `https://doi.org/10.7488/ds/1411`

From ARCOSG to parsing

- We follow the v2 Universal Dependencies guidelines and Lynn *et al.*'s treatment of Irish, with some special cases for Gaelic.
- We assign a UPOS based on the ARCOSG tag, largely following the Irish scheme.
- Some deviations from the PAROLE scheme:
  - Personal prepositions (*agam*, *agad*) count as ADPs.
  - Aspect markers *a'/ag*, *ri*, *gu* are also ADPs (but *cf.* Johannes and Fran's talk).
  - Specials like *airson* 'for' are handled separately.
- Python scripts to lemmatise and assign features.
- In our initial pass we split sentences on full stops and subdivide them more accurately as part of the annotation process.

Universal Dependencies

Requirements set out by Christopher Manning:

- UD needs to be satisfactory on linguistic analysis grounds for individual languages.
- UD needs to be good for linguistic typology, *i.e.*, providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
- UD must be suitable for rapid, consistent annotation by a human annotator.
- UD must be suitable for computer parsing with high accuracy.
- UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing.
- UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, ???).

Universal Dependencies

Some consequences:

- Content words are heads (in PPs the heads are nouns, not the prepositions).
- No distinctions between arguments and modifiers.
- Copula (x is a y) has its own relation. Languages like Russian and Chinese have zero copula.

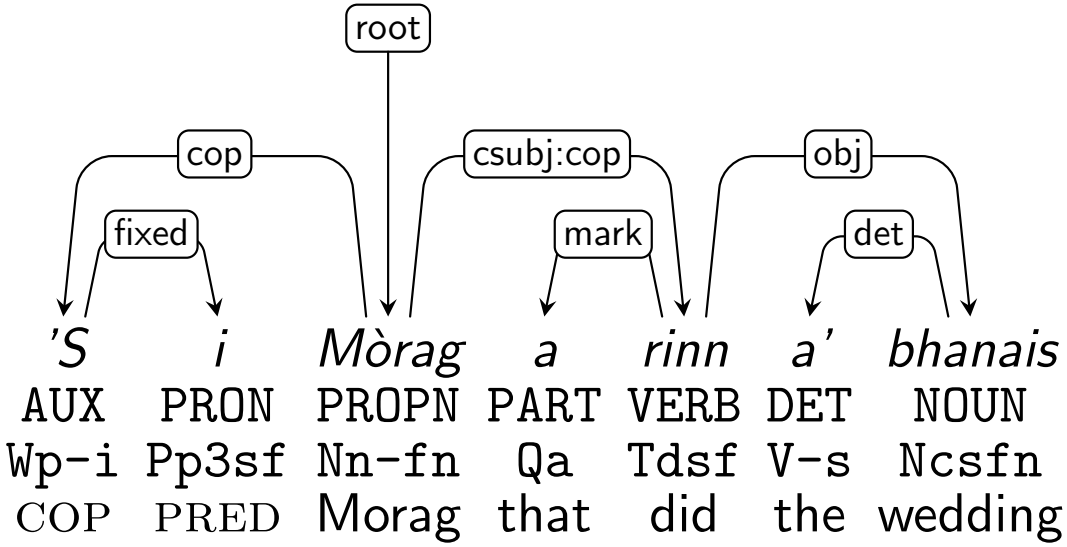
Specific to this work:

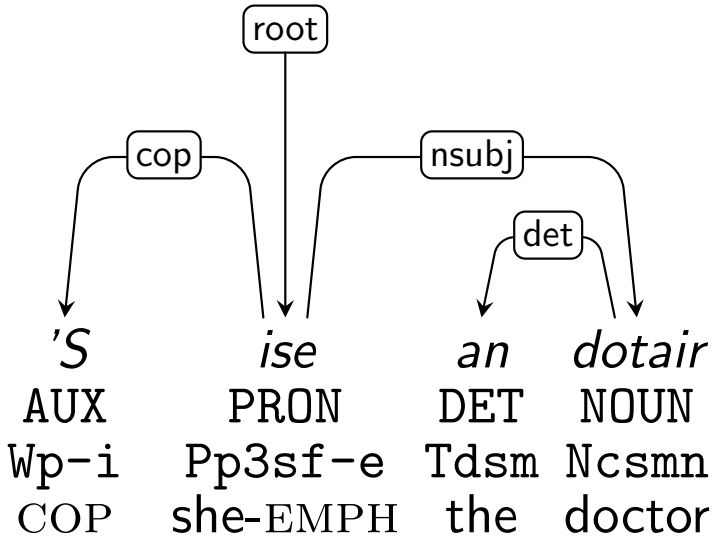
- Treat coordinative phrases like *air sgàth 's gun...* ('on the cause and that...' $\approx$ 'because...'), *fiu 's ged a...* ('worthy and although...' $\approx$ 'even if...'), *còrr is...* ('extra and...' $\approx$ 'more than...') strictly as coordination.
- In NPs ending with a noun in the genitive, say *beagan mhionaidean* 'a few minutes', treat *beagan* as the head and *mhionaidean* as a modifier.

Method

- Convert ARCOSG to CoNLL-U format.
- Put aside sports commentary, conversation, interview and narrative files with no obvious sentence boundaries for now.
- Annotate around a thousand tokens from shortish sentences.
- Train up a parser (initially MaltParser, then udpipe) to annotate the rest.
- Fix errors and iterate, starting with the shortest sentences.

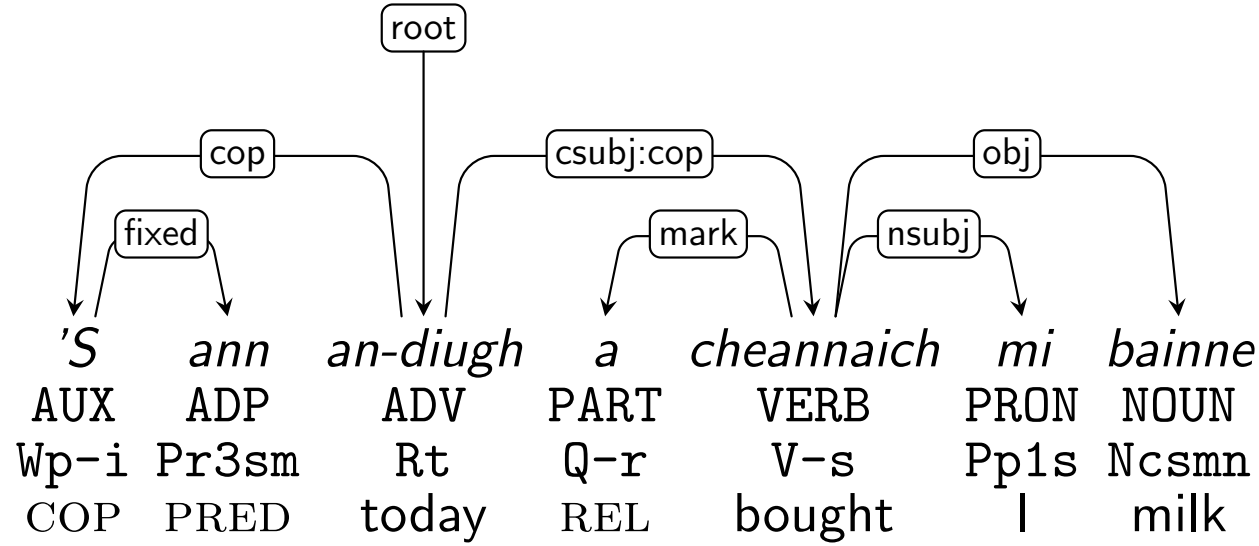
Copula: nouns as roots

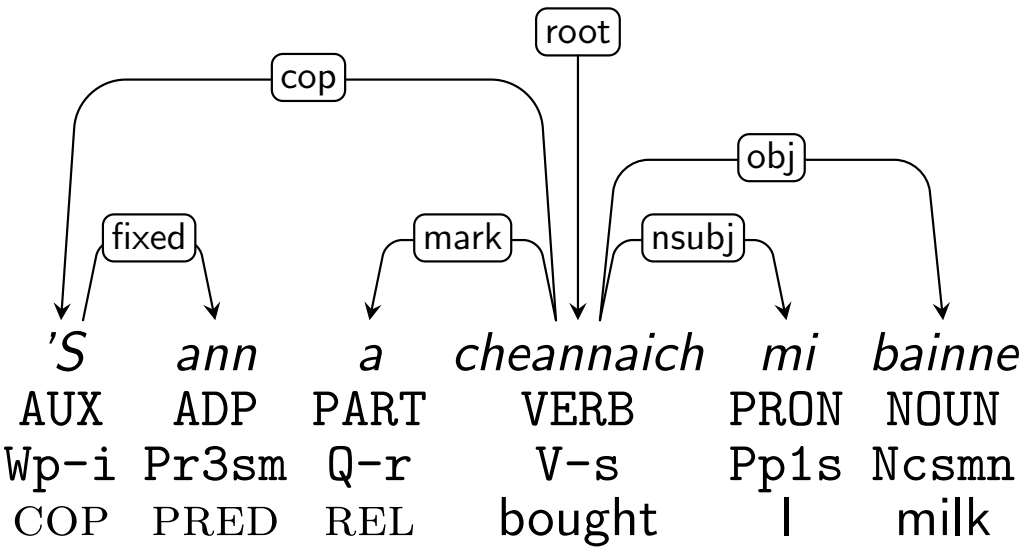




Copula: contrastive use

Examples from Cox (*Geàrr-Ghràmar na Gàidhlig*, 2017):





Results

1021 sentences. 20 031 words. 20 021 tokens.

Transition system	LAS
`projective`	**0.796** [0.752, 0.839]
`swap` (non-projective)	0.789 [0.747, 0.825]
`link2` (non-projective)	0.792 [0.750, 0.835]

Labelled-attachment scores (LAS) for parsing the treebank with different transition systems for udpipe's parsito parser. The LAS are the mean values from ten-fold cross-validation.

Statistics: parts of speech

UPOS	Count	Comments
NOUN	4567	incl. verbal noun
ADP	2711	
PUNCT	2367	
VERB	2029	
PART	1848	
DET	1450	
ADJ	850	
ADV	747	
CCONJ	646	
NUM	243	
AUX	230	*is*, *gur*
SCONJ	218	
X	76	
INTJ	38	

Statistics: Dependency relations

Deprel	Count	Comments
punct	2367	
case	2308	
obl	2145	
nsubj	1761	
mark	1565	
root	1021	
xcomp	832	verbal nouns
nmod	724	
obj	651	
advmod	625	
fixed	547	*'s e*, *an dèidh*
conj	543	
ccomp	529	
amod	517	

Deprel	Count	Comments
flat	390	
xcomp:pred	263	adjectives
cop	226	
acl:relcl	200	
parataxis	195	
nummod	190	
nmod:poss	171	
csubj:cop	111	
appos	89	
compound	82	*fhèin*, *fhìn*
obl:tmod	39	
vocative	15	
case:voc	11	
obl:smod	10	

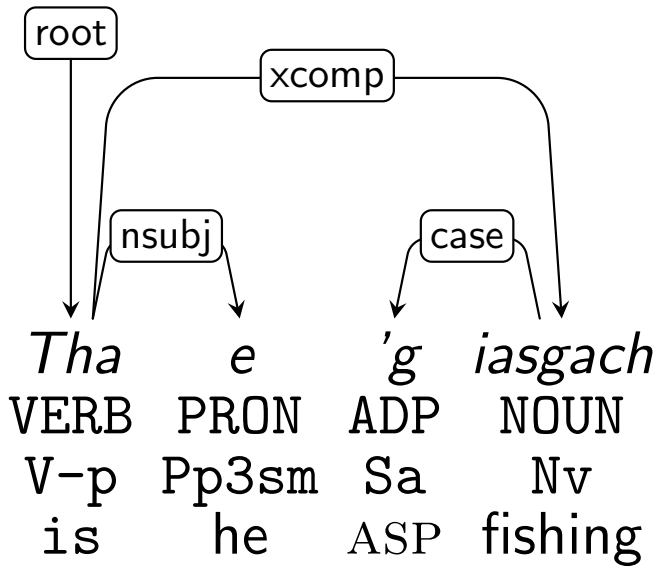
Cosubordination!

Cosubordination is a distinctive feature of Scottish Gaelic found in all registers. This is where *agus*, *is* and *'s* 'and' coordinate a sentence with an incomplete clause where the tense and other features come from the matrix clause (examples from corpus):

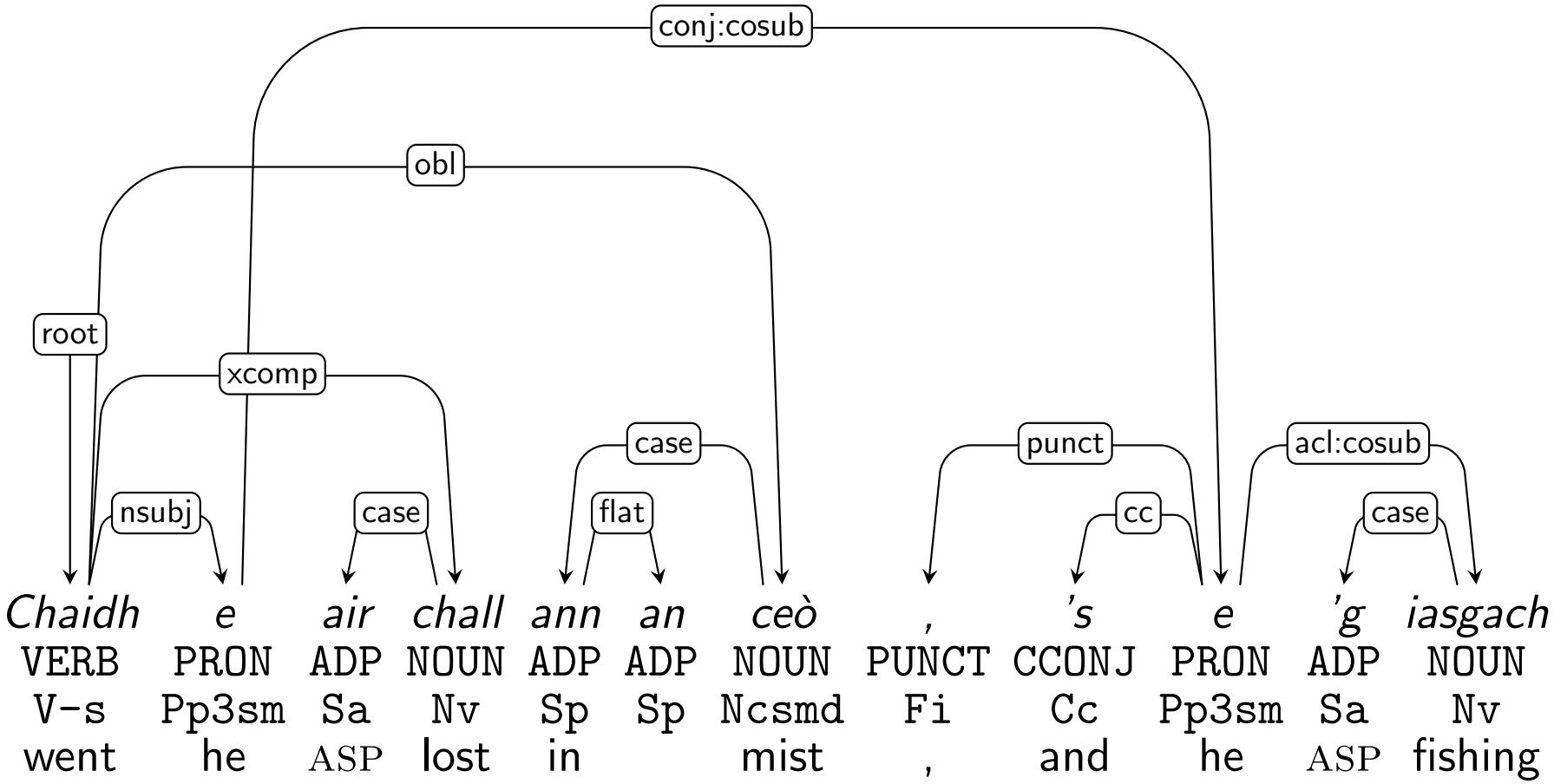
- *Chaidh bratach Bhreatainn a thoirt a-nuas ann an seirbheis taobh muigh an taighe, 's an Last Post ga chluiche.*
- *... tha dà sheòrsa eadar-dhealaichte ann de fhuil A; 'se sin A1 agus A2 is iad fo smachd dà aileal eadar-dhealaichte.*
- *Tha bean-gairm SAND, Lorrainne Mann, ag ràdha gu feuch iad a-nise ri barrachd taic fhaighinn, 's iad a' dol a sgrìobhadh gu comhairlean coimhearsnachd anns an sgìre.*

Cannot be fronted or clefted. Also seen in Irish, where it can be fronted, but couldn't find it in Irish treebank.

He is fishing

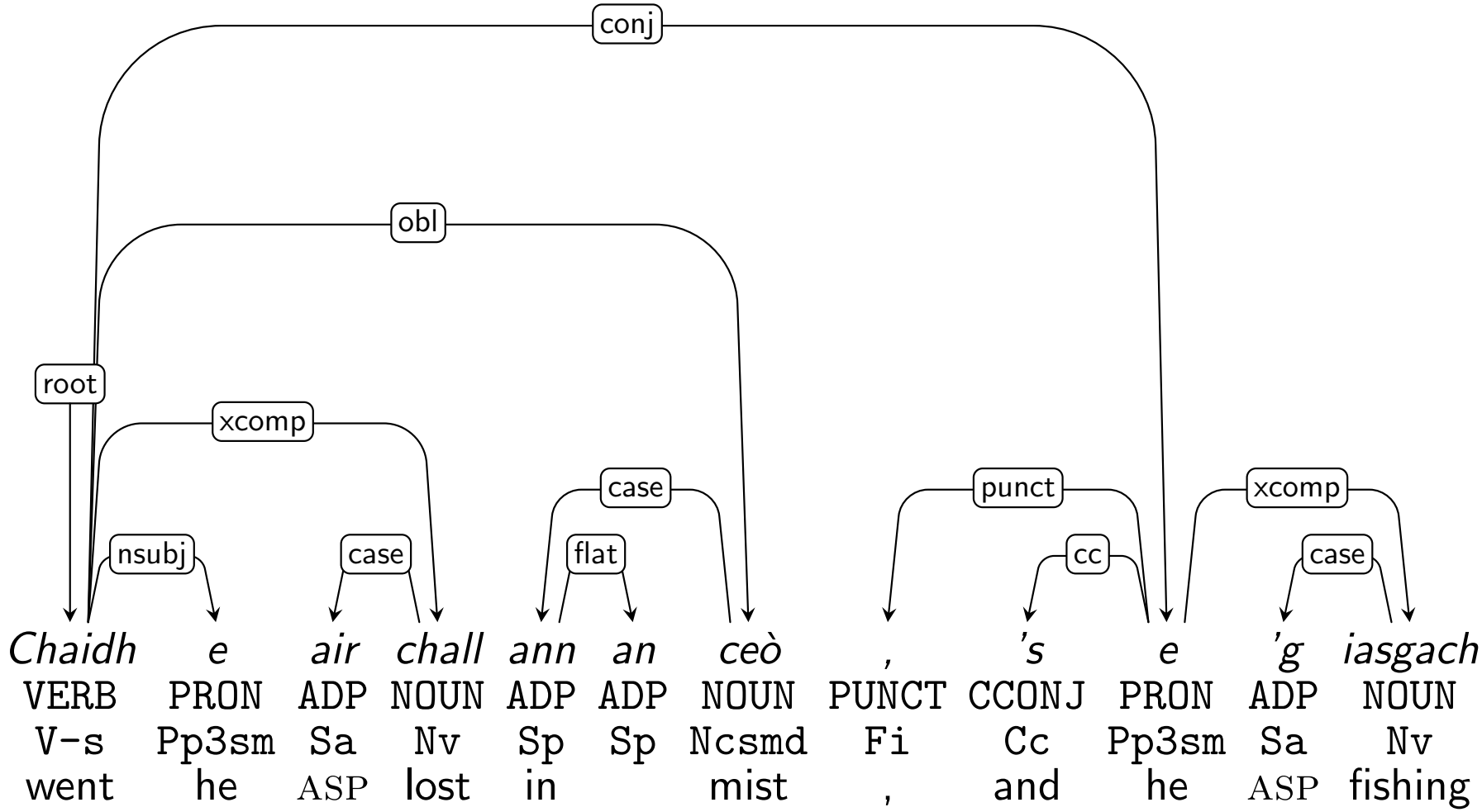


Depictive approach

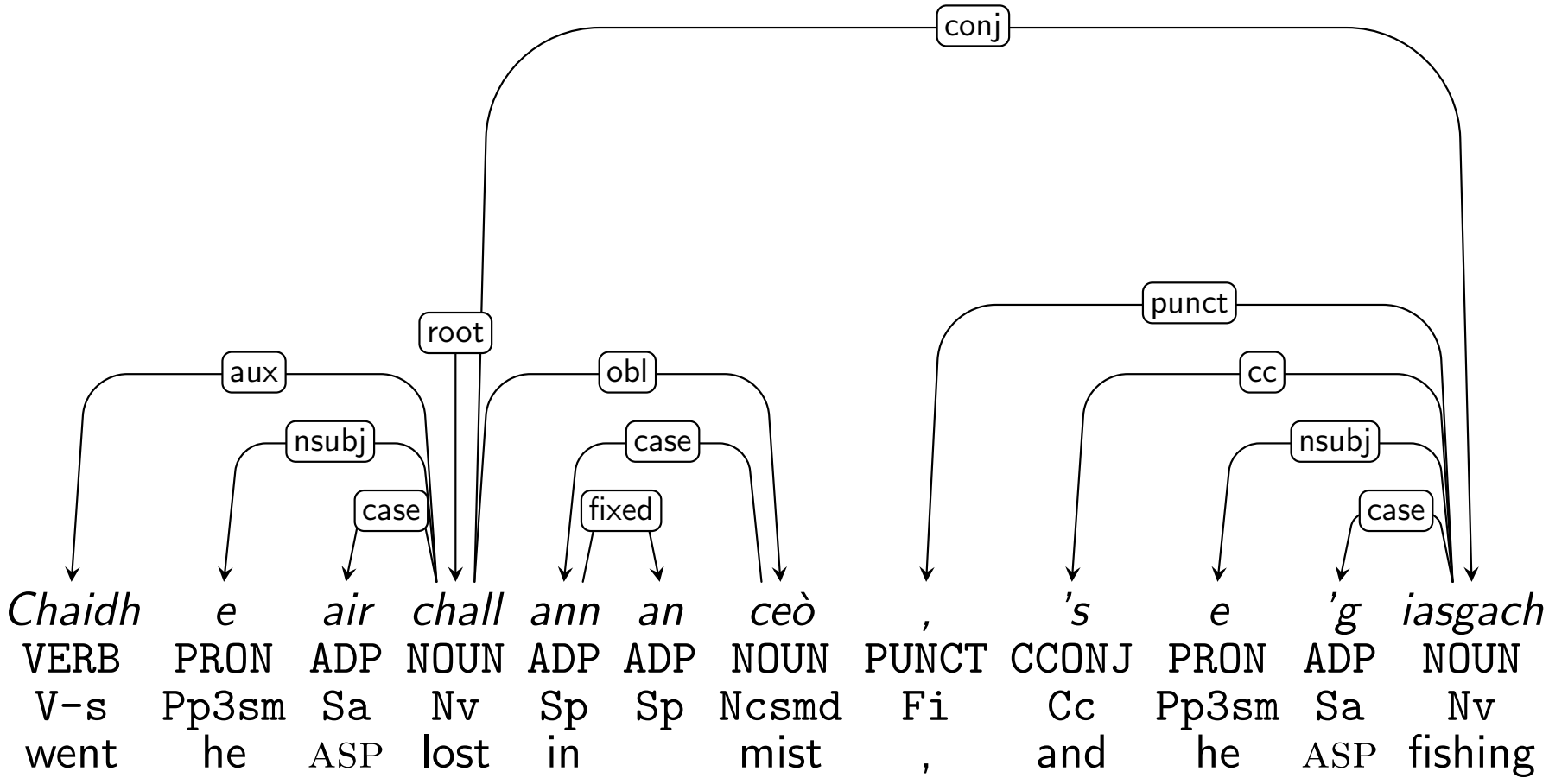


Elliptical approach

(In the paper I used `orphan`, but it should be promoted to `xcomp`. Use `conj:cosub`?)



Auxiliary approach



Comparison

- With *chaidh* and *bi* as heads, we need to choose:
  - The depictive approach coordinates like with like.
  - The elliptical approach is easier to make projective.
  - The elliptical approach implies that there is *something* that has been elided.
- With the verbal noun as head the approach is obvious, but doesn't work for cases with PPs and adjectives.

Conclusions and next steps

- Decide about cosubordination.
- Address conversations.
- Address test/train/dev split. (*qv.* `https://www.aclweb.org/anthology/P19-1267`)
- Write up treatment of comparatives.
- Will go live in v2.5 of Universal Dependencies this November.
- `universaldependencies.org`
- Code (and partly-processed trees) here: `https://github.com/colinbatchelor/gdbank/`
- *Mòran taing dhan a h-uile duine!*
- *Ceist sam bith?*